Санкт-Петербургский Государственный Университет

Биологический факультет

Кафедра генетики

Шиманский Валентин Сергеевич

Оптимизация методов *in silico* генотипирования по аллелям генов главного комплекса гистосовместимости человека

Работа выполнена в лаборатории

«Центр геномной биоинформатики им. Ф.Г. Добржанского»

Научный руководитель: Жернакова Дарья Вячеславовна

Куратор: Даев Евгений Владиславович

Санкт-Петербург

2019

Оглавление

1. Введение

Главный комплекс гистосовместимости человека (ГКГС) или человеческие лейкоцитарные антигены - это комплекс трансмембранных белков, являющийся ключевым в распознавании чужеродных молекул и формировании иммунного ответа. Белки комплекса кодируются более чем 200 генами, которые располагаются на 6 хромосоме. Основная функция ГКГС - презентация антигенов на поверхности клеток для их последующего распознавания Т-лимфоцитами.

Существуют три класса молекул ГКГС. Молекулы ГКГС класса I отвечают за систему распознавания свой – чужой. При помощи молекул класса I на поверхность клеток выносятся аутоантигены, при заражении – вирусные антигены, антигены некоторых внутриклеточных бактерий и опухолевых клеток. Молекулы класса II отвечают за презентацию внеклеточных бактериальных антигенов, токсинов, аллергенов и т.д., и они присутствуют в основном у профессиональных антиген-презентирующих клеток (таких как дендритные клетки и В–лимфоциты). Гены ГКГС III класса кодируют некоторые элементы системы комплемента.

Так как белки ГКГС играю важную роль в иммунной системе, гены, кодирующие эти белки, характеризуются очень высоким полиморфизмом: количество обнаруженных аллелей отдельных генов превышает 10 тысяч. Кроме того, различные аллели входящих в него генов ассоциированы со многими заболеваниями - такими как, например, диабет 2 типа, гепатит В, различные виды рака. Генотипирование по аллелям генов ГКГС проводится для оценки возможности трансплантации органов, так как молекулы данного комплекса определяют гистосовместимость. По этим причинам задача генотипирования образцов по аллелям генов ГКГС является важным этапом диагностики заболеваний, трансплантации органов и поиска новых локусов, сцепленных с заболеваниями.

Из-за множественного аллелизма задача генотипирования аллелей генов ГКГС методами *in silico* сильно осложняется, начиная с этапа секвенирования и выравнивания и заканчивая собственно генотипированием. Разработан ряд программ для *in silico* генотипирования человеческих образцов по аллелям генов ГКГС, но не существует оптимального протокола. Таким образом, для получения качественных результатов необходимо оптимизировать процесс обработки данных и, собственно, генотипирования. В данной работе были протестированы различные методики обработки данных и *in silico* генотипирования образцов полногеномного секвенирования по аллелям генов ГКГС, а также разработан оптимальный алгоритм и методика, позволяющие значительно улучшить качество генотипирования, проводимого в рамках различных исследований. В настоящее

время эта методика применяется для генотипирования образцов проекта "Российские Геномы".

Целью работы является оптимизация методов *in silico* генотипирования образцов полногеномного секвенирования по аллелям генов ГКГС.

Задачи:

1. Подготовка данных полногеномного секвенирования для генотипирования комплекса генов гистосовместимости (выравнивание прочтений и последующее выделение участка 6 хромосомы с исследуемыми генами).
2. Проведение *in silico* генотипирования при помощи специализированных программ.
3. Сравнение работы различных программ типирования и выравнивания, выбор наиболее точного метода.
4. Анализ результатов генотипирования по аллелям генов ГКГС.

2. Обзор литературы

2.1 Механизм функционирования ГКГС

Генная группа комплекса гистосовместимости кодирует группы родственных белков, известных как комплекс лейкоцитарных антигенов человека (ГКГС). Комплекс белков ГКГС помогает иммунной системе отличать собственные белки организма от белков, вырабатываемых чужеродными инвазивными организмами, такими как вирусы и бактерии.

Район локализации генов, кодирующих белки главного комплекса гистосовместимости человека, охватывает 7,6 Мб на хромосоме 6p21, где расположены 252 экспрессируемых локуса, включая несколько ключевых генов иммунного ответа (Horton et al., 2004). Область может быть подразделена на расширенный класс I, классический класс I, классический класс III, классический класс II и расширенный класс II. Для этой области характерна наибольшая степень полиморфизма в геноме человека (Mungal et al., 2003). Большая часть исследований была сосредоточена на роли генов класса II, кодируемых молекул генов *HLA-DR* и *–DQ*, которые представляют экзогенные антигены для распознавания CD4 + Т-хелперными (Th) клетками. Выявлены сильные ассоциации почти со всеми аутоиммунными заболеваниями (АЗ) (Gough et al., 2007).

Функция молекул ГКГС заключается в связывании фрагментов пептидов, полученных от патогенов, и презентации их на клеточной поверхности для распознавания соответствующими Т-клетками. Последствия почти всегда негативны для патогенно-вирусных клеток: активируются макрофаги для уничтожения внутриклеточных бактерий, активируются В-клетки для выработки антител, которые выводят или нейтрализуют внеклеточные патогены. По этой причине идет отбор в пользу любого патогена, мутировавшего таким образом, что он не попадает под действие молекул белков ГКГС.

Два отдельных свойства компонентов ГКГС затрудняют патогенным микроорганизмам уклонение от иммунной реакции. Во-первых, локус генов ГКГС является полигенным: он содержит несколько различных генов класса I и класса II, так что каждый человек обладает набором молекул с различными диапазонами пептид-связывающих свойств. Во-вторых, локус генов ГКГС является высокополиморфным, т.е. существует множество вариантов каждого гена в популяции в целом (Janeway et al., 2001).

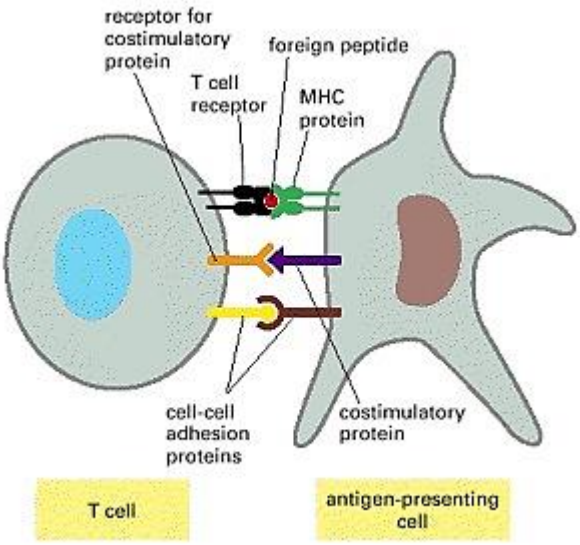

Рисунок 1. Общая схема взаимодействия антиген-презентирующих клеток с Т-клетками посредством ГКГС. (Alberts, 2002)

Молекулы ГКГС класса I экспрессируются почти во всех клетках организма. С их помощью клетки презентируют на поверхность аутоантигены для распознавания Т-киллерами, распознавание осуществляется при помощи рецептора CD8. При заражении клетки вирусом или внутриклеточной бактерией на поверхность наряду с собственными антигенами выносятся чужеродные, что является сигналом для разрушения клетки (Blees et al., 2017).

Презентация антигенных пептидов посредством молекул ГКГС класса I происходит следующим образом. Антигены (собственные или чужеродные) деградируют под воздействием цитозольных и ядерных протеасом. Полученные пептиды переносятся транспортером, связанным с представлением антигенов (TAP), в эндоплазматический ретикулум (ЭР) для доступа к молекулам ГКГС класса I. В ЭР находится гетеродимер ГКГС класса I (Blum et al., 2013). Он состоит из полиморфной тяжелой цепи и инвариантной легкой цепи, состоящей из β-микроглобулина (β2m). Классические гены класса I – *A, B, C* - кодируют различные варианты тяжелой α-цепи.

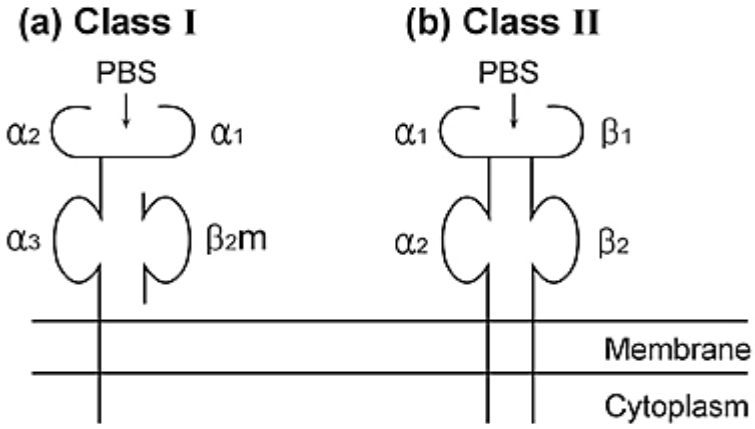

Рисунок 2. Схематичное изображение молекул ГКГС класса I и II. (Choo, 2007)

Пептид — это третий компонент, необходимый для стабильности, так как он проникает глубоко в пептидосвязывающую канавку молекул ГКГС класса I, расположенную между α_1 и α_2 субъединицами. Размер пептида 8-10 аминокислот. Без пептидов молекулы класса I стабилизируются такими белками эндоплазматического ретикулума, как кальретикулин, ERp57 (также известный как PDIA3), дисульфидная изомераза белка (PDI) и специальный тапазин-шаперон. Тапазин взаимодействует с TAP, тем самым связывая транслокацию пептидов в эндоплазматический ретикулум с доставкой пептидов в молекулы ГКГС I класса. Когда пептиды связываются с молекулами класса I, шапероны высвобождаются, и полностью собранные комплексы пептид-молекула ГКГС класса1 покидают эндоплазматический ретикулум для представления на поверхности клетки (Vyas et al., 2008). И наоборот, пептиды и молекулы класса I ГКГС, которые не ассоциируются в эндоплазматическом ретикулуме, возвращаются в цитозоль для деградации (Hughes et al., 1997; Koopmann et al., 2000; Neefjes et al., 2011).

Молекулы ГКГС класса II экспрессируются в профессиональных антиген-презентирующих клетках (дендритных клетках, В-лимфоцитах) и в некоторых непрофессиональных (например, эндотелиоцитах сосудов). Эти молекулы состоят из 2 цепей: α и β. Гены, кодирующие белки ГКГС класса II, попарно кодируют различные варианты цепей (например, ген *DRA* кодирует α-цепь, а ген *DRB* – β-цепь). Многие аллели генов ГКГС класса II являются сильными генетическими маркерами некоторых аутоиммунных заболеваний. По сравнению с изученностью строения и механизмов работы белков ГКГС класса I, способы презентации и влияние полиморфизма на функционирование у класса II изучены менее подробно. Трансмембранные α и β-цепи молекулы ГКГС класса II собираются в ЭР и ассоциируются с инвариантной цепью (Ii). Полученный комплекс Ii-ГКГС II класса транспортируется в поздний эндосомальный пузырек, получивший название пузырек II класса ГКГС (MIIC). Здесь Ii переваривается,

оставляя остаточный пептид Ii (CLIP) в пептидной канавке гетеродимера ГКГС класса II (находится между α_1 и β_1 субъединицами). В MIIC молекулам ГКГС класса II нужен HLA-DM (H2DM у мышей) для облегчения замены CLIP-фрагмента на конкретный пептид, полученный из белка, разложившегося по эндосомальному пути. Затем молекулы класса II ГКГС секретируются из клетки, чтобы представить свой пептидный груз клеткам CD4+T. В клетках экспрессируется модификация HLA-DM под названием HLA-DO (H2O у мышей), который ассоциируется с HLA-DM и ограничивает его активность по отношению к более кислотным компартментам, модулируя тем самым связывание пептидов с молекулами класса II ГКГС (Neefjes et al., 2011; Denzin et al., 2005).

2.2 История открытия и изучения ГКГС

Изучение белков ГКГС человека началось в 60-70 гг. XX века. Первый антиген ГКГС был обнаружен в 1958 г. французским исследователем Ж. Доссетом в сыворотке крови пациентов после переливания методом лейкоагглютинации (Dausset, 1958). Изначально этот антиген был назван MAC, впоследствии переименован в *HLA-A2*. Спустя 6 лет, в 1964 г. состоялось первое международное собрание по вопросам гистосовместимости, организованное Бернардом Амосом. Основной целью этого собрания была демонстрация и обмен методами обнаружения составляющих ГКГС. Были представлены: метод лейкоагглютинации, непрямой тест на потребление антиглобулина Коломбани, тест на фиксацию комплемента Уилсоном, тест смешанной агглютинации Мильгромом и Метцгаром и метод микро-лимфоцитотоксичности Теразаки.

На втором собрании, организованном Джоном Ван Родом, одинаковые наборы из 45 клеток анализировались в 14 лабораториях с использованием их собственных методов для сравнения особенностей, обнаруженных независимо. Таким образом MAC, изначально обнаруженный Доссетом, был независимо получен в 5 лабораториях методами лейкоагглютинации, фиксации комплемента, клеточной цитотоксичности.

В 1967 г. состоялось собрание Всемирной Организации Здравоохранения (ВОЗ), в ходе которой было принято решение об учреждении комитета по созданию единой номенклатуры для исследований ГКГС. Первое заседание комитета состоялось в 1968 г. Гены комплекса гистосовместимости были переименованы в антигены лейкоцитов человека (*HLA*), а также разделены на два класса, I и II.

Большой взрыв в открытии новых антигенов произошел перед четвертой международной конференцией, посвящённой ГКГС. Метод микро-лимфоцитотоксического теста был выбран основным для последующего изучения антигенов. Начиная с 4-ой конференции все последующие, в основном, заключались в обмене образцами сыворотки

и аналогичных испытаниях в разных лабораториях до очередного ежегодного подведения итогов (Thorsby, 2009).

Более 50 лет изучается тесная связь между генами ГКГС и аутоиммунными заболеваниями (АЗ). Ассоциация компонентов гаплотипа *HLA-DRB1-DQA1-DQB1* была обнаружена для многих АЗ, включая ревматоидный артрит, диабет 1 типа, аутоиммунный гепатит. Молекулы, кодируемые этой областью, играют ключевую роль в презентации экзогенного антигена CD4 + Th-клеткам, что подчёркивает важность этого пути в инициации и прогрессировании АЗ. Хотя другие компоненты областей ГКГС класса I и III также были исследованы на предмет ассоциации с АЗ, кроме ассоциации *HLA-B*27* с болезнью Бехтерева, было трудно определить дополнительные локусы, независимые от неравновесного состояния сильных связей (СВ) с генами гены, кодирующими молекулы ГКГС класса II. Недавние успехи в статистическом анализе СВ и наборе больших объемов данных по АЗ позволили повторно исследовать гены, кодирующие молекулы ГКГС класса I и III. Ассоциация области ГКГС класса I, независимая от известных эффектов ГКГС класса II, в настоящее время обнаружена для нескольких АЗ, включая сильную связь *HLA-B* с диабетом типа 1 и *HLA-C* с рассеянным склерозом и болезнью Грейвса. Эти результаты являются дополнительным доказательством возможной роли бактериальной или вирусной инфекции и CD8 + Т-клеток в начале развития АЗ. Достижения в определении первичных ассоциаций в регионе генов, кодирующих молекулы ГКГС, с АЗ не только улучшат понимание механизмов, стоящих за патогенезом заболевания, но также могут помочь в разработке новых терапевтических методов в будущем (Gough et al., 2007).

2.3 Молекулярно-генетические методы типирования.

Долгое время основным способом генотипирования по аллелям генов ГКГС были серологические методы, основанные на лимфоцитотоксических тестах. В таких методах для определения генотипа используется сыворотки с наборами антител для разных антигенов ГКГС. Если антитела распознают соответствующие антигены, то при добавлении комплемента происходит клеточный лизис, который определяется с помощью люминесцентной микроскопии (Althaf et al. 2017). Серологические методы имеют ряд серьёзных недостатков, таких как низкая точность и чувствительность, а также высокая трудоёмкость. Поэтому в настоящее время чаще применяются генетические методы генотипирования, что значительно улучшило возможности и точность генотипирования в сравнении с ранее использовавшимися методами.

В данном обзоре рассмотрены следующие наиболее типовые подходы к генотипированию по аллелям генов ГКГС: протоколы основанных на полимеразной

цепной реакции (ПЦР), использующие специфические праймерные последовательности (SSP) или специфичные олигонуклеотидные зонды (SSO) и методы типирования на основе секвенирования (SBT) (Nowak et al., 2012).

2.3.1 Генотипирование с помощью олигонуклеотидных зондов

Этот метод генотипирования состоит из трёх основных этапов. Первый этап - амплификация фрагментов, соответствующих генам ГКГС, с помощью ПЦР. Первые применения ПЦР с использованием последовательности специфичных олигонуклеотидов (PCR-SSO) в области антигена лейкоцитов человека (HLA) начались с локуса *DQA1*, затем были расширены на класс I и другие сайты II класса. Чтобы устранить неоднозначности, были введены протоколы, в которых используются ПЦР-праймеры, предназначенные для амплификации всей гипервариабельной области конкретного локуса ГКГС (группоспецифичных праймеров). Второй этап - инкубация полученных продуктов ПЦР в присутствии меченых олигонуклеотидов, предназначенных для обнаружения различных полиморфных положений, специфичных для аллеля или аллельной группы. Праймеры для локусов *HLA-A*, *-B* и *-C* обычно дают локус-специфический продукт, охватывающий экзоны 2 и 3. Праймер для *HLA-DR* дает продукт из экзона 2 (Dunckley 2012). Третий этап - визуализация результатов. Продукты гибридизации можно проследить либо с помощью колориметрической (стрептавидин – биотин), рентгеновской (дигоксигенин – CSPD) или флуоресцентной (FITC, PE) систем обнаружения. Для достижения более быстрой, надежной, автоматизированной техники типирования были объединены методики обнаружения при помощи микрогранул и флуоресценции и внедрены в эту область (технология XMAP ™). Для каждого локуса используется серия микросфер, различаемых по их специфическому цвету, происходящему из двух внутренних флуоресцентных красителей. Каждая микросфера связана с одним зондом, который способен гибридизоваться с меченным биотином комплиментраным ампликоном. Как только происходит гибридизация, ее можно количественно определить, измеряя флуоресцентный сигнал, исходящий от захваченных (стрептавидин-PE) ампликонов. В настоящее время существуют две коммерчески доступные системы, которые отличаются по шкале зондов и методам, используемым для амплификации и денатурации (Dalva et al., 2014).

2.3.2 Праймер-специфичное генотипирование

Данный метод использует специализированные праймеры. Пары праймеров подобраны таким образом, чтобы они гибридизовались только на определенную аллель или семейство аллелей генов ГКГС. После ПЦР амплифицированные фрагменты ДНК

разделяются при помощи электрофореза в агарозном геле, визуализируются при помощи бромистого эттидия или SYBR green, под воздействием ультрафиолета, затем полученные результаты интерпретируются. Интерпретация основывается на присутствии или отсутствии специфичных ПЦР продуктах в реакционных пробирках с известными аллель-специфичными праймерами. В связи с обширным полиморфизмом локуса ГКГС, наборы для данного метода поставляются только в коммерческом виде в формате наборов микропробирок для ПЦР с предварительно размещенными специализированными праймерами с точно описанной характеристикой аллели и расположением микропробирок (Dunckley 2012, Dalva et al., 2014).

2.3.3 Генотипирование, основанное на секвенировании.

Этот подход состоит из двух основных этапов: ПЦР амплификации и секвенирования полученных продуктов. 2 – 4 экзоны генов ГКГС являются наиболее полиморфными среди других фрагментов генов. В данном методе для амплификации обычно используют смеси праймеров, специфичных к локусу/экзону. Последующее секвенирование также выполняется при помощи локус/экзон – специфичных праймеров и несущих цветовую метку ПЦР дидезоксинуклеотиды. Для секвенирования обычно используется секвенирование нового поколения или метод обрыва цепи Сэнгера (Dalva et al., 2014).

2.4 Методы генотипирования in silico

Несмотря на быстрый прогресс геномных технологий и на всё большую доступность генотипирования и секвенирования для исследователей и врачей, определение аллелей генов ГКГС затруднено из-за значительного сходства последовательностей внутри кластера и исключительно высокой изменчивости локусов. Поэтому применяемые молекулярные подходы основаны на специфичных методах амплификации локуса ГКГС и секвенирования, сопряженных с дополнительными затратами и дополнительными сроками выполнения работ. В связи с этим возникла потребность в разработке биоинформатических методов генотипирования по аллелям генов ГКГС, опирающихся только на результаты полногеномного (или полноэкзомного) генотипирования или секвенирования. Удобство таких методов состоит в том, что, например, для геномных исследований и поиска ассоциаций с заболеваниями можно использовать стандартные данные генотипирования, без необходимости проведения дополнительных молекулярных тестов.

2.4.1 Методы генотипирования по данным микрочипов

Эта группа методов основана на использовании специализированных ДНК микрочипов, содержащих последовательности с определёнными однонуклеотидными заменами, комплементарными исследуемым участкам. Образец наносится на чип для гибридизации с зондами, и микрочип сканируется для распознавания гибридизации. Так как генотипы образцов влияют на гибридизацию с зондами, соответствующими данному локусу, сканирование позволяет определить генотипы образцов. Особенность этого метода в том, что генотипирование проводится только для заранее заданных зондами позиций. Большинство микрочипов содержат олигонуклеотидные зонды в том числе и к определённым позициям локуса ГКГС. Однако из-за особенностей генов локуса ГКГС определение аллелей этих генов на основании только отдельных однонуклеотидных замен (HLA imputation) представляет собой сложную задачу, для решения которой был разработан ряд программных пакетов, таких как SNP2HLA(Jia et al., 2013) и HLA*IMP (Dilthey et al. 2011). Эти программы, как правило, используют референсную базу данных аллелей генов ГКГС для большого количества образцов и, на основании этой базы данных, воссоздают аллели и гаплотипы генотипируемых образцов с помощью анализа неравновесного сцепления, подходов, основанных на графах, и других методов.

2.4.2 Методы генотипирования по данным секвенирования

Основным недостатком генотипирования по данным микрочипов является очень ограниченный исходный объём информации, так как микрочипы покрывают только небольшое количество замен в генах ГКГС, что приводит к невысокой точности генотипирования. Развитие геномных технологий привело к появлению эффективных методов секвенирования, стоимость которых с каждым годом снижается. Эти методы позволяют узнать всю последовательность генома в локусе ГКГС, что позволяет значительно улучшить точность генотипирования, в том числе для редких аллелей.

На данный момент разработан ряд программ для *in silico* генотипирования по аллелям генов ГКГС на основании данных полногеномного или полноэкзомного секвенирования. В основе работы этих программ, как правило, лежит алгоритм линейного выравнивания прочтений на референсную базу данных аллелей, однако есть и исключения, использующие, например, выравнивание на популяционный граф. Программы, используемые в данной работе, были выбраны на основании нескольких сравнительных исследований, таких как (Kiyotani et al., 2017). Ниже мы даём краткое описание алгоритма программ, используемых в данной работе.

2.4.4.1 Athlates

Программа Athlates использует базу данных аллелей генов ГКГС IMGT/HLA. Данные секвенирования, относящиеся к экзонам ГКГС, выравниваются на имеющуюся базу данных. Программа отдельно учитывает случаи, когда прочтение или пара прочтений выравнивается неоднозначно (больше чем на один ген локуса). Результат выравнивания записывается в сжатом формате BAM, из которого извлекаются прочтения, выровненные с целевым геном ГКГС (например, HLA-A), и в специальном BED файле регистрируются все аллели этого гена. Аналогичным образом извлекаются прочтения, выровненные на нецелевые гены. Далее происходит сборка. На этапе сборки используются прочтения/пары прочтений, однозначно выровненные с геном-мишенью. Парные прочтения, выровненные на одну и ту же референсную аллель, объединяются в один при помощи следующего метода. Парное прочтение (r_0, r_1), выровненное к одной аллели разделяется на подстроки (r^p_0, o_0, r^s_1, o_1), где r^p_0 и r^s_1 — это не перекрывающиеся префикс и суффикс r_0 и r_1 соответственно, а o_0 и o_1 — перекрывающиеся подстроки. o_0 и o_1 должны быть одинаковой длины. Если o_0 и o_1 не являются пустыми строками, из них формируется общая строка o_m. Если эти строки являются пустыми, вместо них подставляются «N».

Затем каждое считывание инициализируется как контиг с базовой частотой, записываемой по каждой позиции. Прочтения делятся на подстроки (l-меры), длина которых изначально равна длине самого прочтения, а затем пошагово уменьшается до минимального значения в 40 нуклеотидов. Контиги, делящие наиболее длинные l-меры, с большей вероятностью относятся к одному гаплотипу, поэтому для каждого значения l контиги представляются как набор l-меров, сортированных по убыванию частоты.

Следующий этап генотипирования – составление списка кандидатных аллелей. Каждая аллель целевого гена разделяется на экзоны, для каждого из которых подбираются наиболее подходящие контиги («совпадение»). Качество каждого «совпадения» определяется длинной и схожестью подстрок контига с последовательностью экзона. Для описания степени схожести используется расстояние Хэмминга. Максимально допустимое расстояние Хэмминга равно 2, варианты с большим значением не участвуют в анализе. Затем считается суммарное расстояние Хэмминга для всех «совпадений» по проверяемому аллелю. Идеальным вариантом является нулевое суммарное расстояние Хэмминга. Однако истинный аллель может иметь ненулевое расстояние по нескольким причинам: во-первых, мелкие экзоны и часть длинных экзонов могли быть не охвачены при экзомном секвенировании, и, во-вторых, у генотипируемого образца может быть новый аллель с небольшим количеством мутаций по сравнению с известным аллелем. Поэтому

целесообразно рассматривать все аллели в пределах некоторого допустимого диапазона расстояний (по умолчанию, 2).

На основании данных по покрытию аллели прочтениями формируется итоговая пара (или несколько пар в случае недостатка данных для однозначного выявления) определенных аллелей, поступающая на вывод программы (Liu et al., 2013).

2.4.2.2 HLA-LA

Программа HLA-LA использует иной подход к генотипированию образцов по аллелям генов ГКГС. В качестве референса используется эталонный граф популяции (PRG). PRG — это ориентированный граф, в котором альтернативные аллели, вставки и делеции представлены как альтернативные пути через граф, и в котором ортологические и идентичные области свертываются локально для моделирования потенциальной рекомбинации. PRG строится таким образом, чтобы охватить весь спектр ГКГС, гены гаплотипов и экзонов для 46 генов (в основном, HLA). Используемый алгоритм короткого чтения следует классической парадигме "seed-and-extend", т.е. определяет области точной идентичности между графом и картируемым прочтением, и пытается расширить их с помощью динамического программирования, позволяя отслеживать несовпадения, инсерции и делеции. Каждое выравнивание следует за возможным шагом через PRG граф. Важно то, что PRG граф кодирует информацию об изменениях последовательности, и алгоритм отображения использует эту информацию, что обеспечивает точное выравнивание при наличии однородных вариантов и более точное количественное определение неоднозначности отображения. Наконец, при условии, что прочтения будут сопоставлены с PRG и базой данных о возможных базовых гаплотипах (т.е. последовательностях, аллельных ГКГС), делается вывод о наиболее вероятной паре базовых гаплотипов и качественной оценке с использованием простой вероятностной структуры (Dilthey et al., 2016).

3. Материал и методы

Материалом данной работы являются результаты полногеномного секвенирования нового поколения 7 образцов крови, описанных ранее (Zhernakova et al., 2018). Образцы полногеномного секвенирования, использованные в данной работе, представляют из себя 5 образцов крови пациентов, больных аутоиммунным гепатитом, и двух здоровых родителей одного из пациентов. Кровь была отправлена в три центра секвенирования: СПбГУ (секвенатор HiSeq-4000, Ресурсный Центр, Петергоф), Illumina (секвенатор HiSeq-X10, Великобритания) и Macrogen (секвенатор HiSeq-X10, Южная Корея). Таким образом, были получены 3 повторности (технических репликата) секвенирования. Последующие этапы анализа данных проводились для каждой повторности.

Разработка алгоритма проводилась поэтапно. На первом этапе по литературным данным (Kiyotani et al., 2017; Dilthey et al., 2016; Liu eta al., 2013) был подобран набор программ, показывавших наиболее точные по сравнению с эталонными молекулярными методами результаты генотипирования. Для сравнения были выбраны два метода, имеющие принципиально разные алгоритмы генотипирования: Athlates (Liu eta al., 2013) и HLA-LA (Dilthey et al., 2016).

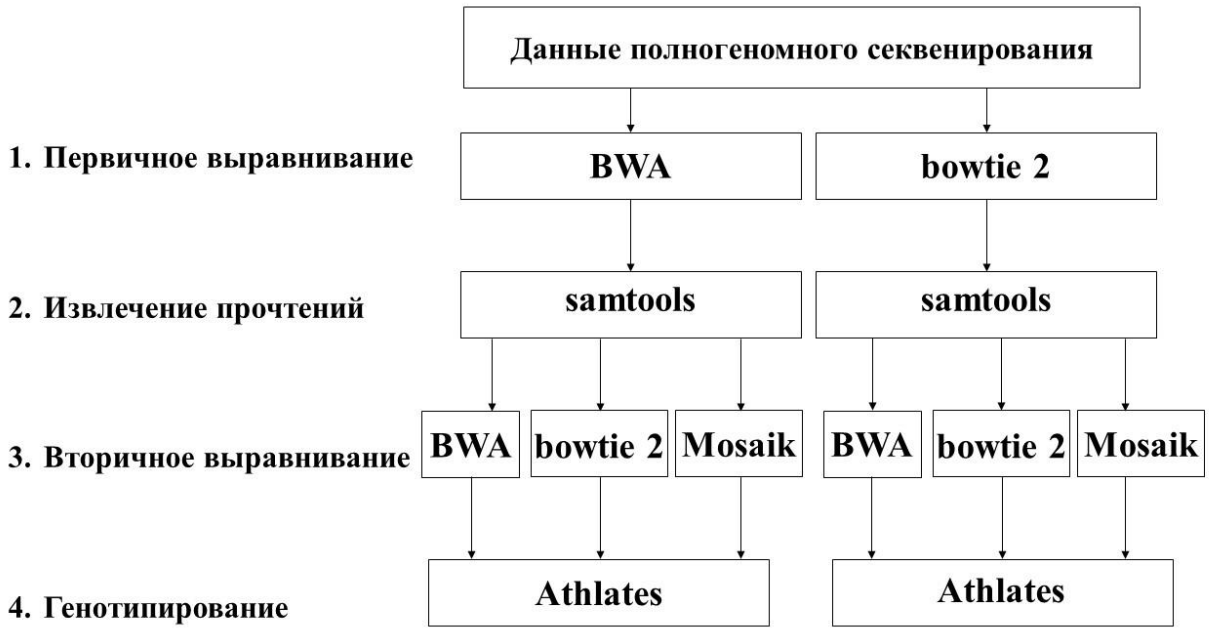

Рисунок 1. Общая схема тестирования программ подготовки данных генотипирования результатов при помощи программы Athlates

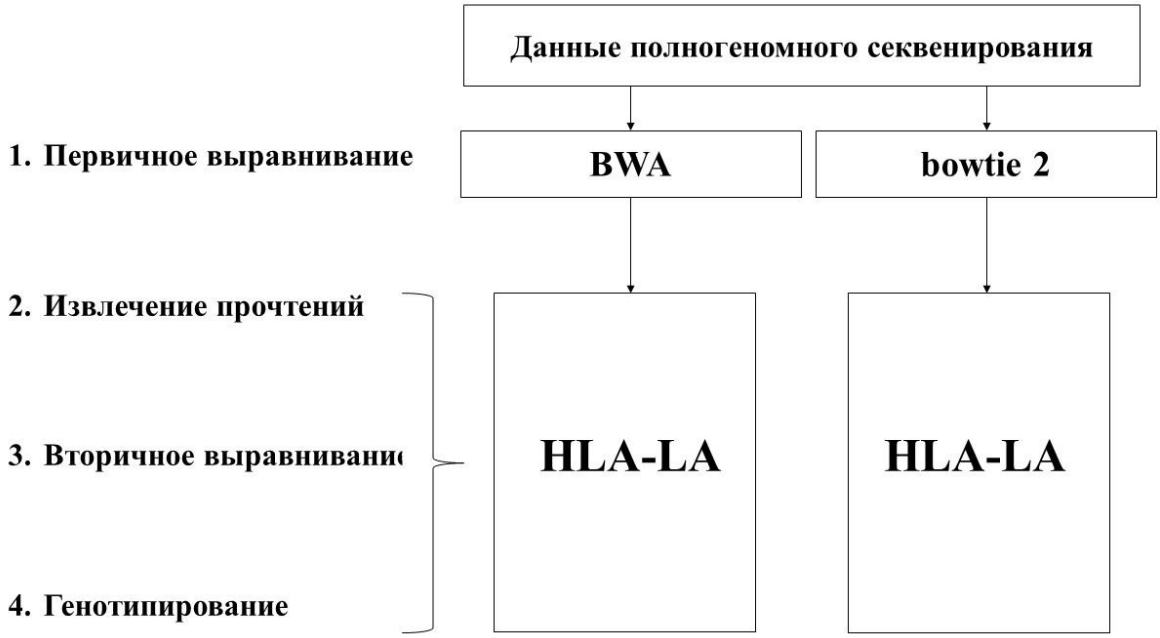

Рисунок 2. Общая схема тестирования программ подготовки данных генотипирования результатов при помощи программы HLA-LA

Подготовка данных для генотипирования по ГКГС и само генотипирование может проводиться различными методами. Использование разных комбинаций этих методов приводят к отличиям в полученных генотипах. На втором этапе были протестированы различные комбинации методов подготовки данных и генотипирования в целях определения наиболее точного. Структурная схема представлена на Рис.1 и 2. Точность определялась путем сравнения полученных генотипов с эталоном. В качестве эталона использовались данные молекулярного типирования 7 образцов проекта по аллелям генов ГКГС, проведенного по методике, описанной в статье Tang et al., 2012. Оценка точности каждого метода генотипирования *in silico* проводилась путем подсчета процента правильно определенных аллелей генов (где правильными генотипами считались результаты эталонного молекулярного метода) из всех генотипированных аллелей данного гена по всем семи образцам. На основании точности генотипирования подбирался оптимальный набор программ подготовки данных и, собственно, генотипирования.

Сравнение результатов проводилось без статистической оценки значимости различий. Это объясняется тем, что вероятность правильного генотипирования разная для разных аллелей, и, таким образом, невозможно статистически оценить значимость различия количества правильных генотипов для всего локуса или для каждого гена с использованием стандартных методов. Для статистической оценки сравнения результатов для каждого аллеля в отдельности необходимо значительно большая выборка – в случае 7 образцов этого сделать невозможно из-за высокого количества возможных аллелей.

Поэтому, в результатах исследования приводится процент ошибок генотипирования для каждого метода, которые сравниваются без статистической оценки достоверности различий. Дальнейшее исследование будет включать выборку большего объема и разработку статистической методологии сравнения методов.

Результаты секвенирования (парные прочтения длиной 150-151 п.о.) обрабатывались следующим образом. После контроля качества, проведенного с помощью FastQC (Andrew S., 2010), прочтения были выровнены на референсный геном человека (версия GRCh38) с помощью bowtie2 (Langmead et al., 2012) и BWA (Li et al., 2009). Затем выровненные прочтения использовались для генотипирования по аллелям генов ГКГС. Для исследования влияния увеличения покрытия на результаты генотипирования также анализировались объединенные данные центров Macrogen и Illumina (объединение проводилось после полногеномного выравнивания).

Для данного исследования были выбраны две программы генотипирования по аллелям генов ГКГС – Athlates (Liu et al., 2013) и HLA-LA (Dilthey et al., 2016), которые используют принципиально разные алгоритмы. Так как эти программы генотипирования принимают на вход данные разного типа, сравнение методов подготовки данных для этих двух программ отличалось. Были проведены следующие исследования, причем на каждом этапе выбиралась оптимальная программа, которая и использовалась в следующих:

1. Athlates
   1.1. Сравнение программ для первичного полногеномного выравнивания (этап 1 на рис. 1)
   1.2. Сравнение программ для вторичного выравнивания на базу данных генов ГКГС (этап 3 на рис. 1)
2. HLA-LA
   2.1. Сравнение программ для первичного полногеномного выравнивания (этап 1 на рис. 2)
3. Сравнение программ генотипирования по ГКГС (этап 4 на рис. 2)

Так как каждый образец был отсеквенирован 3 раза, сравнения были проведены по трем повторностям.

3.1 Подготовка данных и генотипирование программой Athlates.

Из выровненных данных полногеномного секвенирования были выделены прочтения, относящиеся к региону генов комплекса гистосовместимости человека (chr6:28000000-34000000) при помощи программы samtools (Li et al., 2009) (стандартные параметры запуска); далее эти прочтения были заново выровнены на базу данных генов

комплекса гистосовместимости, предоставляемой программой Athlates (Liu et al., 2013) (стандартные параметры запуска), при помощи bowtie2 (Langmead et al., 2012) (стандартные параметры запуска), BWA (Li et al., 2009) (стандартные параметры запуска) и Mosaik (Lee et al., 2014) (параметры запуска, рекомендованные Athlates). Такое вторичное выравнивание повышает точность дальнейшего анализа. Далее аллели генов ГКГС были генотипированы при помощи программы Athlates (Liu et al., 2013). При помощи программы samtools и базы данных IMGT/HLA (Li et al., 2009) общий набор данных разбивается по отдельным генам, программа Athlates анализирует прочтения, находит лучшие совпадения по базе данных, формирует пары аллелей, и затем определяет наиболее вероятных кандидатов.

3.2 Подготовка данных и генотипирование программой HLA-LA.

Второй программой, которую мы использовали для генотипирования, является HLA-LA (Dilthey et al., 2016). HLA-LA выполняет типирование на основе эталонного графа популяции (PRG), и использует метод линейной проекции для согласования с графом. Для ее работы не требуются шаги извлечения прочтений, относящихся к исследуемому локусу, и вторичное выравнивание, достаточно результатов полногеномного секвенирования. Для итоговой оценки использовались только те аллели, значение вероятности правильного определения для которых превышала 95% (согласно собственному значению программы, рассчитываемому в ходе генотипирования).

4. Результаты и обсуждение

В данной работе было проведено генотипирование 7 образцов полногеномного секвенирования по генам ГКГС с помощью различных комбинаций методов (см. Рис. 1 и 2) и сделан выбор оптимальной комбинации на основании точности генотипирования. Некоторые из комбинаций оказались невозможны, по причине несоответствия входных/выходных данных или несоответствия алгоритмов. Такие случаи мы также описали в результатах (Таблица 1).

Таблица 1. Программы, используемые при разработке алгоритма, «+» отмечены работающие сочетания, «-» отмечены случаи, когда программы оказались несовместимы. Для программы HLA-LA вторичное выравнивание не требуется, поэтому соответствующие ячейки пусты

Программы генотипирования	Программы выравнивания				
	Первичное		Вторичное		
	BWA	bowtie2	BWA	bowtie2	Mosaik
Athlates	-	+	+	-	+
HLA-LA	+	+			

Далее приведены результаты сравнения комбинаций методов подготовки данных отдельно для двух программ генотипирования по ГКГС: Athlates и HLA-LA. Далее следует сравнение этих двух программ генотипирования при подготовке данных выбранным на первом этапе оптимальным методом.

4.1 Генотипирование с помощью программы Athlates

Было проведено сравнение результатов при использовании разных программ для выравнивания прочтений на базу данных генов ГКГС: BWA, bowtie2 и Mosaik (Lee et al., 2014). На данных, вторично выровненных при помощи bowtie2, программу Athlates запустить не удалось. По этой причине было проведено генотипирование с помощью Athlates на результатах вторичного выравнивания, выполненного с помощью BWA и Mosaik. Сравнительный график процента совпадений генотипов с молекулярным типированием при использовании этих двух программ приведен ниже (рисунок 5).

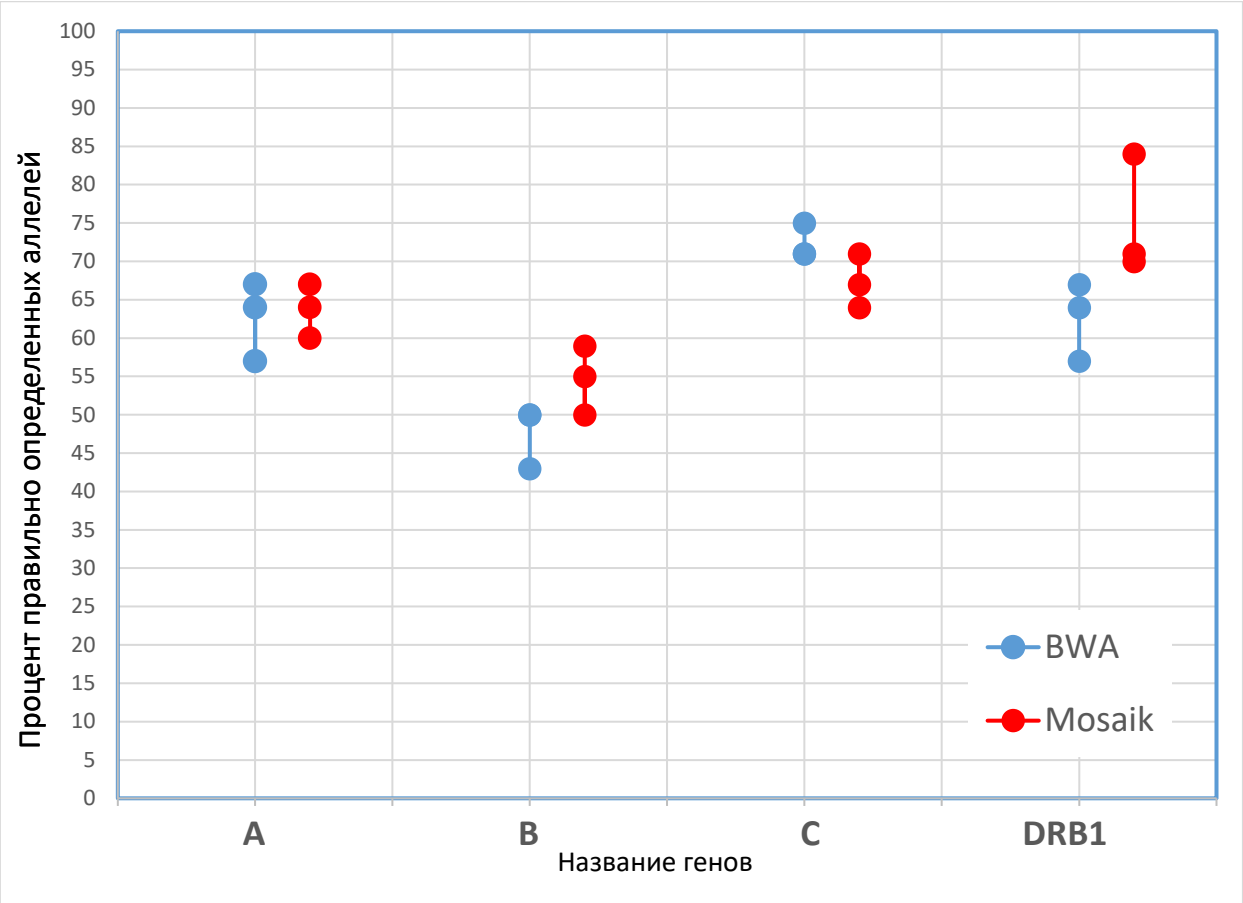

Рисунок 5. Сравнительные результаты генотипирования по генам ГКГС программой Athlates, где вторичное выравнивания выполнено при помощи программ BWA и Mosaik. По оси X идут названия генов, по оси Y – соответствующие им проценты правильно генотипированных аллелей (совпавших с эталоном) сочетанием программы Athlates с BWA (синий цвет) или с Mosaik (красный цвет). Каждая точка соответствует точности генотипирования, посчитанной по 7 образцам. Линиями соединены три повторности секвенирования

По данным, представленным на рис.5, можно сделать предположение, что различия между использованием BWA и Mosaik для вторичного выравнивания не существенны. Данные, полученные при выборе Mosaik для вторичного выравнивания, совпадают с эталоном в среднем приблизительно на 64%, при выборе BWA – на 61%, однако необходимо провести сравнение на большей выборке для получения более достоверного результата. Для дальнейшего исследования программа Athlates запускалась на данных, полученных с помощью программы BWA.

4.2 Генотипирование с помощью программы HLA-LA

Следующим этапом работы стало тестирование программы HLA-LA (Dilthey et al., 2016). Анализировались результаты полногеномного секвенирования, так как все этапы обработки, такие как извлечение прочтений, относящихся к локусу ГКГС и дальнейшее вторичное выравнивание входят в состав алгоритма работы программы. Первоначальное

(полногеномное) выравнивание было выполнено с помощью программ bowtie2 и BWA. Результаты сравнения генотипирования приведены на Рис. 6.

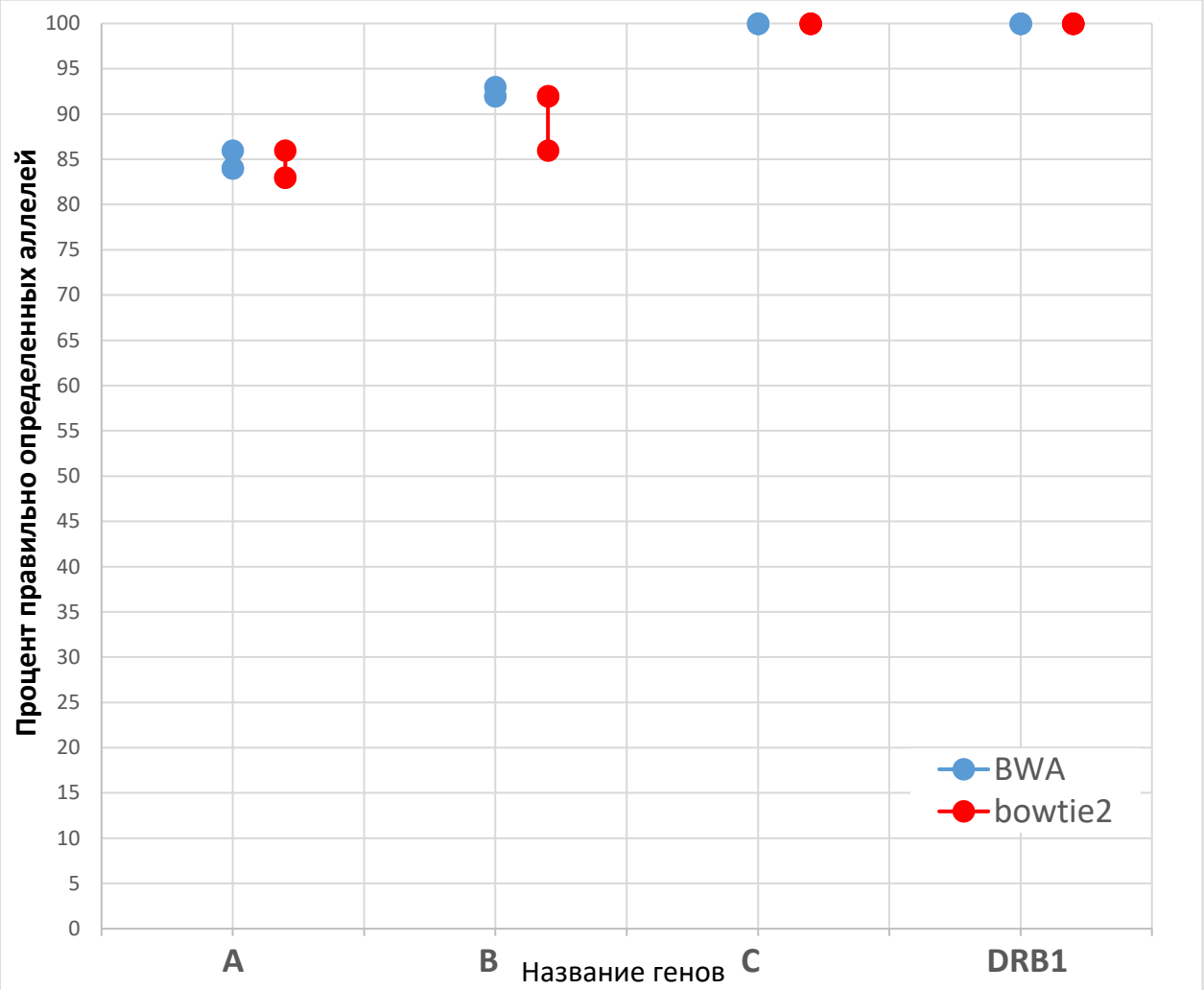

Рисунок 6. Результаты сравнения точности генотипирования с помощью программы HLA-LA для 7 образцов, где первичное выравнивания выполнено при помощи программ BWA и bowtie2. По оси X идут названия генов, по оси Y – соответствующие им проценты правильно генотипированных аллелей (совпавших с эталоном) сочетанием программы HLA-LA с BWA (синий цвет) или с bowtie2 (красный цвет). Каждая точка соответствует точности генотипирования, посчитанной по 7 образцам. Линиями соединены три повторности секвенирования. Когда точность одинакова для нескольких повторностей, для них показана одна точка

По данным, представленным на рисунке 6, можно сделать предположение, что различия между использованием BWA и bowtie2 в качестве программы первичного выравнивания, не существенны. Данные генотипирования приведены в таблице 2 приложения. Процент совпадений с эталоном для программы HLA-LA составляет в среднем примерно 95, что превосходит показатели точности программы Athlates. Также, по сравнению с программой Athlates, HLA-LA способна определять аллели значительно большего числа генов. Помимо основного набора генов *A*, *B*, *C*, *DRB1* также представлены аллели генов *DQA1*, *DQB1*, *DPA1*, *DRB3*, *DRB4*, *E*, *F*, *G*.

Несмотря на то, что различия между сочетаниями программ BWA+HLA-LA и bowtie2+HLA-LA выявлены не были, для дальнейшей работы мы использовали BWA в качестве программы выравнивания, так как именно она обычно используется для выравнивания полногеномных данных для последующего полногеномного генотипирования.

4.3 Сравнение программ генотипирования по аллелям генов ГКГС

В результате сравнений различных комбинаций программ подготовки данных для генотипирования, были выбраны следующие комбинации: bowtie2 → BWA → Athlates и BWA → HLA-LA.

Ниже представлен сравнительный график результатов работы этих двух комбинаций (рисунок 7).

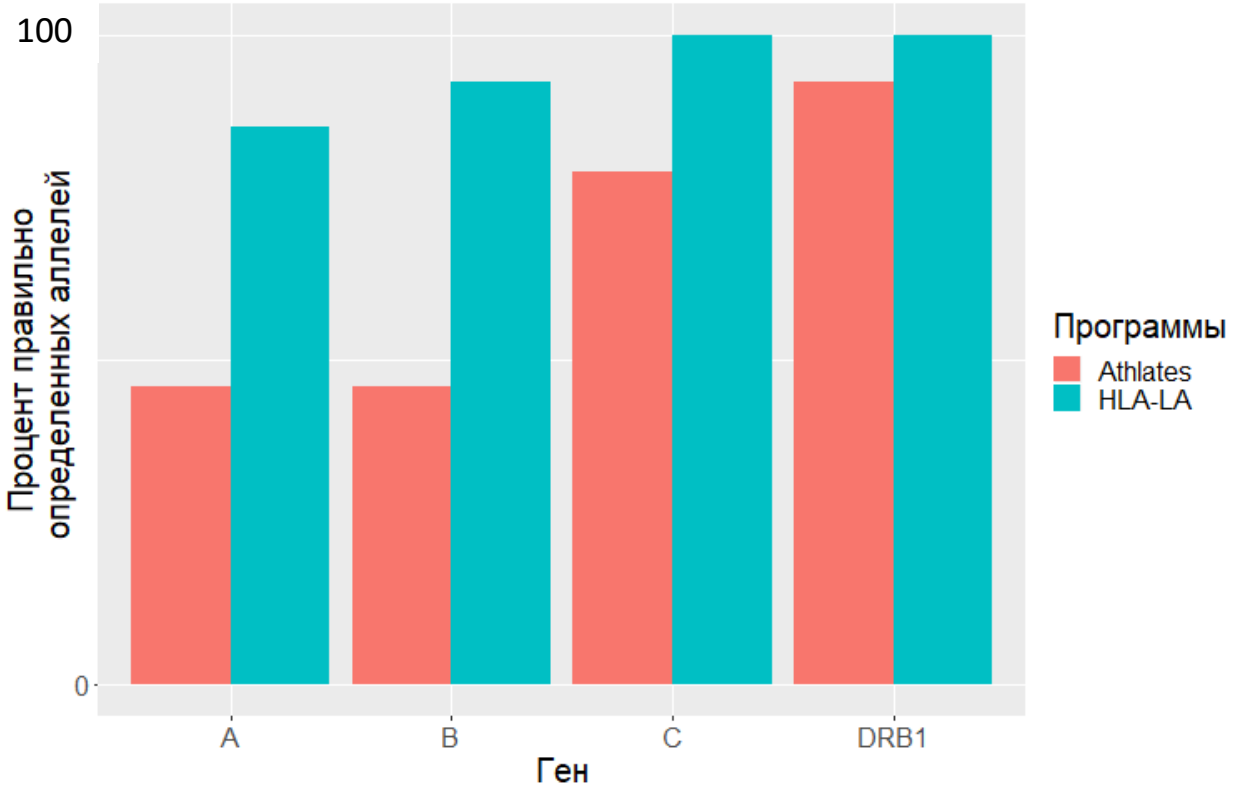

Рисунок 7. Сравнительные результаты генотипирования 7 образцов с помощью программ генотипирования HLA-LA (синие столбцы) и Athlates (красные столбцы). По оси X указаны названия генов, по оси Y – соответствующие им проценты совпадающих с эталоном генотипированных аллелей

Как видно из графика, HLA-LA показывает более высокую долю правильно генотипированных аллелей, чем Athlates. Однако, разница в результатах может объясняться как большей точностью программы генотипирования (HLA-LA по сравнению с Athlates), так и программой первичного выравнивания (BWA по сравнению с bowtie2, так как комбинация BWA → BWA → Athlates не работала).

Как было замечено, оценка статистической значимости разницы в точности генотипирования осложняется тем, что вероятность ошибки определения разных аллелей разная. Чтобы получить представление о приблизительной значимости разницы, мы также оценили эту значимость, пренебрегая разницей в ошибке определения. Использование критерия Фишера показало, что различия в точности генотипирования значимы для генов *A* и *B* (p-значения 0.0461 и 0.0128 соответственно) и не значимы для генов *C* и *DRB* (p-значения > 0.95).

Так как данные, используемом для исследования, включали 5 пациентов, больных аутоиммунным гепатитом, также были проанализированы результаты генотипирования с точки зрения аллелей, сцепленных с заболеванием. Небольшой размер выборки и отсутствие информации по частотам аллелей здоровых людей из той же популяции не позволил нам найти новые аллели, сцепленные с аутоиммунным гепатитом. Однако мы проверили частоту встречаемости аллелей, которые по литературным данным увеличивают риск развития аутоиммунного гепатита. У 5 пациентов были обнаружены аллели гена *DRB1*, а именно DRB1*07:01 и DRB1*13:01, ассоциированные с аутоиммунным гепатитом (Gought et al., 2007., Tawandee et al., 2006), причем частота этих аллелей выше, чем у здоровых людей, что указывает на возможную причину увеличения риска развития заболевания у этих пациентов. Однако для получения достоверных результатов необходимо увеличение выборки.

5. **Заключение**

В результате данной работы была проведена оценка двух методов in silico генотипирования образцов по аллелям генов ГКГС, а именно сравнение результатов работы программ Athlates и HLA-LA с результатами молекулярного типирования. Также было проведено сравнение результатов генотипирования в зависимости от выбора программ для первичного и вторичного выравнивания прочтений на базу данных генов ГКГС. Из полученных результатов следует, что генотипирование главного комплекса гистосовместимости человека программными методами на основе данных полногеномного секвенирования является сложной проблемой и требует тщательного тестирования для получения достоверных результатов.

В ходе представленной работы:

1. Были выявлены совместимые и несовместимые (BWA на этапе 1 + Athlates, bowtie2 на этапе 3 + Athlates) комбинации программ.
2. По результатам генотипирования можно предположить, что использование различных программ для вторичного выравнивания не влияет на качество генотипирования
3. Найдено наиболее оптимальное и точное сочетание программ BWA для выравнивания результатов полногеномного секвенирования и HLA-LA для определения аллелей генов ГКГС. Это сочетание программ было выбрано по следующим причинам:

- самая высокая точность генотипирования среди всех проверяемых комбинаций методов;
- доступна информация о степени достоверности генотипов, вычисляемая HLA-LA, которая может использоваться для более или менее строгой фильтрации результатов;
- на вход программе подаются результаты полногеномного выравнивания, то есть нет необходимости в дополнительных этапах обработки данных;
- программа активно разрабатывается и усовершенствуется.

Эти программы будут использоваться для генотипирования образцов по ГКГС в текущих и планируемых проектах, таких как «Российские Геномы» и полноэкзомный анализ ассоциаций инфицирования вирусом гепатита B.

Список литературы

1. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. T Cells and MHC Proteins. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26926/
2. Blees, A., Januliene, D., Hofmann, T., Koller, N., Schmidt, C., Trowitzsch, S., … Tampé, R. (2017). Structure of the human MHC-I peptide-loading complex. Nature, 551(7681), 525–528. https://doi.org/10.1038/nature24627
3. Blum, J. S., Wearsch, P. A., & Cresswell, P. (2013). Pathways of antigen processing. Annual Review of Immunology, 31, 443–473. https://doi.org/10.1146/annurev-immunol-032712-095910
4. Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics, 81(5), 1084–1097. https://doi.org/10.1086/521987
5. Charles A Janeway, J., Travers, P., Walport, M., & Shlomchik, M. J. (2001). The major histocompatibility complex and its functions. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK27156/
6. Choo, S. Y. (2007). The HLA system: genetics, immunology, clinical testing, and clinical implications. Yonsei Medical Journal, 48(1), 11–23. https://doi.org/10.3349/ymj.2007.48.1.11
7. Dalva, K., & Beksac, M. (2014). HLA Typing with Sequence-Specific Oligonucleotide Primed PCR (PCR-SSO) and Use of the LuminexTM Technology. In Methods in molecular biology (Clifton, N.J.) (Vol. 1109, pp. 87–99). https://doi.org/10.1007/978-1-4614-9437-9_6
8. Dausset, J. (1958). Iso-leuco-anticorps. Acta Haematologica, 20(1–4), 156–166. https://doi.org/10.1159/000205478
9. Dilthey, A. T., Moutsianas, L., Leslie, S., & McVean, G. (2011). HLA*IMP—an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics, 27(7), 968–972. https://doi.org/10.1093/bioinformatics/btr061
10. Dilthey, A. T., Moutsianas, L., Leslie, S., & McVean, G. (2011). HLA*IMP—an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics, 27(7), 968–972. https://doi.org/10.1093/bioinformatics/btr061
11. Dilthey, A. T., Moutsianas, L., Leslie, S., & McVean, G. (2011). HLA*IMP—an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics, 27(7), 968–972. https://doi.org/10.1093/bioinformatics/btr061
12. Dilthey, A. T., Gourraud, P.-A., Mentzer, A. J., Cereb, N., Iqbal, Z., & McVean, G. (2016). High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. PLoS Computational Biology, 12(10), e1005151. https://doi.org/10.1371/journal.pcbi.1005151
13. Dunckley, H. (2012). HLA Typing by SSO and SSP Methods. https://doi.org/10.1007/978-1-61779-842-9_2
14. Dunckley, H. (2012). HLA Typing by SSO and SSP Methods. https://doi.org/10.1007/978-1-61779-842-9_2
15. Dunckley, H. (2012). HLA Typing by SSO and SSP Methods. In Methods in molecular biology (Clifton, N.J.) (Vol. 882, pp. 9–25). https://doi.org/10.1007/978-1-61779-842-9_2

16. Gough, S. C. L., & Simmonds, M. J. (2007). The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. Current Genomics, 8(7), 453–465. https://doi.org/10.2174/138920207783591690

17. Horton, R., Wilming, L., Rand, V., Lovering, R. C., Bruford, E. A., Khodiyar, V. K., … Beck, S. (2004). Gene map of the extended human MHC. Nature Reviews Genetics, 5(12), 889–899. https://doi.org/10.1038/nrg1489

18. Hughes, E. A., Hammond, C., & Cresswell, P. (1997). Misfolded major histocompatibility complex class I heavy chains are translocated into the cytoplasm and degraded by the proteasome. Proceedings of the National Academy of Sciences, 94(5), 1896–1901. https://doi.org/10.1073/pnas.94.5.1896

19. Jia, X., Han, B., Onengut-Gumuscu, S., Chen, W.-M., Concannon, P. J., Rich, S. S., … de Bakker, P. I. W. (2013). Imputing amino acid polymorphisms in human leukocyte antigens. PloS One, 8(6), e64683. https://doi.org/10.1371/journal.pone.0064683

20. Karnes, J. H., Shaffer, C. M., Bastarache, L., Gaudieri, S., Glazer, A. M., Steiner, H. E., … Roden, D. M. (2017). Comparison of HLA allelic imputation programs. PloS One, 12(2), e0172444. https://doi.org/10.1371/journal.pone.0172444

21. Kiyotani, K., Mai, T. H., & Nakamura, Y. (2017). Comparison of exome-based HLA class I genotyping tools: identification of platform-specific genotyping errors. Journal of Human Genetics, 62(3), 397–405. https://doi.org/10.1038/jhg.2016.141

22. Koopmann, J. O., Albring, J., Hüter, E., Bulbuc, N., Spee, P., Neefjes, J., … Momburg, F. (2000). Export of antigenic peptides from the endoplasmic reticulum intersects with retrograde protein translocation through the Sec61p channel. Immunity, 13(1), 117–127. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10933400

23. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. https://doi.org/10.1038/nmeth.1923

24. Lee, W.-P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P., & Marth, G. T. (2014). MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. PLoS ONE, 9(3), e90581. https://doi.org/10.1371/JOURNAL.PONE.0090581

25. Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B. J., Waite, D., & Davey, R. P. (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Frontiers in Genetics, 4, 288. https://doi.org/10.3389/fgene.2013.00288

26. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

27. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

28. Liu, C., Yang, X., Duffy, B., Mohanakumar, T., Mitra, R. D., Zody, M. C., & Pfeifer, J. D. (2013). ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. Nucleic Acids Research, 41(14), e142. https://doi.org/10.1093/nar/gkt481

29. Ma, X., & Qiu, D. K. (2001). Relationship between autoimmune hepatitis and HLA-DR4 and DRbeta allelic sequences in the third hypervariable region in Chinese. World Journal of Gastroenterology, 7(5), 718–721. https://doi.org/10.3748/WJG.V7.I5.718

30. Mungall, A. J., Palmer, S. A., Sims, S. K., Edwards, C. A., Ashurst, J. L., Wilming, L., … Beck, S. (2003). The DNA sequence and analysis of human chromosome 6. Nature, 425(6960), 805–811. https://doi.org/10.1038/nature02055

31. Neefjes, J., Jongsma, M. L. M., Paul, P., & Bakke, O. (2011). Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nature Reviews Immunology, 11(12), 823–836. https://doi.org/10.1038/nri3084

32. Nowak, J., Mika-Witkowska, R., & Graczyk-Pol, E. (2012). Genetic Methods of HLA Typing. https://doi.org/10.1007/978-3-642-29467-9_21

33. Oliveira, L. C., Porta, G., Marin, M. L. C., Bittencourt, P. L., Kalil, J., & Goldberg, A. C. (2011). Autoimmune hepatitis, HLA and extended haplotypes. Autoimmunity Reviews, 10(4), 189–193. https://doi.org/10.1016/j.autrev.2010.09.024

34. Park, I., & Terasaki, P. (2000). Origins of the first HLA specificities. Human Immunology, 61(3), 185–189.

35. Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., & Kohlbacher, O. (2014). OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics, 30(23), 3310–3316. https://doi.org/10.1093/bioinformatics/btu548

36. Tanwandee, T., Wanichapol, S., Vejbaesya, S., Chainuvati, S., & Chotiyaputta, W. (2006). Association between HLA class II alleles and autoimmune hepatitis type 1 in Thai patients. Journal of the Medical Association of Thailand = Chotmaihet Thangphaet, 89 Suppl 5, S73-8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17722299

37. Thorsby, E. (2009). A short history of HLA. Tissue Antigens, 74(2), 101–116. https://doi.org/10.1111/j.1399-0039.2009.01291.x

38. Vyas, J. M., Van der Veen, A. G., & Ploegh, H. L. (2008). The known unknowns of antigen processing and presentation. Nature Reviews Immunology, 8(8), 607–618. https://doi.org/10.1038/nri2368

39. Zheng, X., Shen, J., Cox, C., Wakefield, J. C., Ehm, M. G., Nelson, M. R., & Weir, B. S. (2014). HIBAG—HLA genotype imputation with attribute bagging. The Pharmacogenomics Journal, 14(2), 192–200. https://doi.org/10.1038/tpj.2013.18

40. Zhernakova, D. V., Kliver, S., Cherkasov, N., Tamazian, G., Rotkevich, M., Krasheninnikova, K., … O'Brien, S. J. (2018). Analytical "bake-off" of whole genome sequencing quality for the genome Russia project using a small cohort for autoimmune hepatitis. PLoS ONE, 13(7). https://doi.org/ы10.1371/journal.pone.0200423

Приложение

Таблица 1. Результаты генотипирования ГКГС программой Athlates. Для каждого образца даны генотипы аллелей генов ГКГС, полученные в каждой повторности (с помощью молекулярного типирования или из полногеномного секвенирования в трех массивах данных. Красным выделены аллели, отличающиеся от молекулярного типирования

Образец	**ген ГКГС**	**Центры секвенирования**								
		Молекулярное типирование		**Peterhof-HiSeq4000**		**Illumina-X10**		**Macrogen-X10**		**Illumina+Macrogen**
пациент-трио1	A	02:01	02:132	02:01:01	02:01:01	02:01:01	02:01:01	02:01:01	02:01:01	02:01:01
	B	44:03	45:01:01	45:01	45:01	44:03:01	44:15	44:03:01	44:15	45:01
	C	16:01:01	16:01:01	16:01:01	16:01:01	16:01:01	16:01:01	16:01:01	16:01:01	16:01:01
	DRB1	07:01:01	12:01:01	12:01:01	12:01:01	12:07	07:01:01	12:01:01	12:01:01	07:01:01
отец-трио1	A	02:01	02:132	02:01:01	02:01:01	02:01:01	02:01:01	02:01:01	02:01:01	02:01:01
	B	44:02:01	45:01:01	44	44:02:01:02S	44:02:01:02S	45:01	44:09	44:02:01:02S	45:01
	C	05:01	16:01:01	16:02:01	08:02:01	16:01:01	05:01:01	16:01:01	05:01:01	16:01:01
	DRB1	10:01:01	12:01:01	12:01:01	12:01:01	10:01:01	12:01:01	10:01:01	12:01:01	10:01:01
мать-трио1	A	02:01	03:01:01	02:01:01	03:01:01	02:01:01	03:01:01	02:01:01	03:01:01	02:01:01
	B	39:06:02	44:03	44:03:01	39:06:02	44:03:01	44:03:01	44:03:01	39:06:02	44:03:01
	C	07:02	16:01:01	16:01:01	07:02:01	16:01:01	07:02:01	16:01:01	07:02:04	16:01:01
	DRB1	07:01:01	08:01:01	08:01:03	08:01:01	08:01:03	08:01:01	NA	NA	08:01:03
Пациент2	A	31:01:02	68:01:02	31:01:02	68:01:02	31:01:02	31:01:02	31:01:02	31:01:02	31:01:02
	B	13:01	51:01:01	51:01:01	51:01:01	NA	NA	13:01:01	51:01:01	51:01:01
	C	04:03:01	14:02:01	14:02:01	04:03	14:02:01	14:02:01	14:02:01	04:03	14:02:01
	DRB1	13:01:01	15:01:01	13:01:01	15:01:01	15:01:01	13:01:01	13:01:01	15:01:01	13:01:01

Образец	ген ГКГС	Центры секвенирования								
		Молекулярное типирование		Peterhof-HiSeq4000		Illumina-X10		Macrogen-X10		Illumina+Macrogen
Пациент3	A	02:01	33:03:01	02:01:01	33:03:01	33:03:01	02:217:02	02:01:01	33:03:01	02:01:01
	B	53:01:01	53:01:01	35:01:01	53:01:01	35:01:01	35:01:01	35:01:01	53:01:01	35:01:01
	C	04:01:01	07:02	07:02:01	04:28	07:02:01	04:01:01	07:02:01	04:01:01	07:02:01
	DRB1	11:01:02	15:01:01	11:01:02	15:01:01	11:01:02	11:01:02	11:01:02	15:01:01	11:01:02
Пациент4	A	24:02	68:68	24:40N	24:40N	24	24	24:02:01	24:02:01	24:02:01
	B	07:02	40:02:01	07:02:01	40:98	40:98	40:02:01	07:02:01	40:02:01	07:02:01
	C	07:01	15:02:01	15:02:01	07	15:02:01	07:18	07:18	15:02:01	07:18
	DRB1	04:05:01	15:03:01	15:03:01	15:03:01	15:03:01	15:03:01	15:03:01	15:03:01	04:05:01
Пациент5	A	02:01	26:01:01	-	-	02:01:01	26:01:01	?	?	02:01:01
	B	08:01:01	39:01	-	-	42:01:01	39:10:01	39:01:01	39:01:01	39:04
	C	07:01	12:03:01	-	-	07:01:01	12:03:01	07:01:01	12:23	07:01:01
	DRB1	03:01:01	13:01:01	-	-	13:01:01	13	03:01:01	03:01:01	03:01:01

Образец	Гены	Молекулярное типирование		центры секвенирования					
				Macrogen		Peterhof		Illumina	
Пациент 1	A	02:01	02:132	02:01	02:01	02:01	02:01	02:01	02:01
	B	44:03	45:01	44:03	45:01	44:03	45:01	44:03	45:01
	C	16:01	16:01	16:01	16:01	16:01	16:01	16:01	16:01
	DRB1	07:01	12:01	07:01	12:01	07:01	12:01	07:01	12:01
Пациент 2	A	02:01	03:01	02:01	03:01	02:01	03:01	02:01	03:01
	B	39:06	44:03	39:06	44:03	39:06	44:03	39:06	44:03
	C	07:02	16:01	07:02	16:01	07:02	16:01	07:02	16:01
	DRB1	07:01	08:01	07:01	08:01	07:01	08:01	07:01	08:01
Пациент 3	A	02:01	02:132	02:01	02:01	02:01	02:01	02:01	02:01
	B	44:02	45:01	45:01	44:02	45:01	44:02	44:02	45:01
	C	05:01	16:01	05:01	16:01	05:01	16:01	05:01	16:01
	DRB1	10:01	12:01	10:01	12:01	10:01	12:01	10:01	12:01
Пациент 4	A	02:01	33:03	02:01	33:03	02:01	33:03	02:01	33:03
	B	53:01	53:01	53:01	35:01	35:01	53:01	35:01	53:01
	C	04:01	07:02	04:01	07:02	04:01	07:02	04:01	07:02
	DRB1	11:01	15:01	11:01	15:01	11:01	15:01	11:01	15:01
Пациент 5	A	31:01	68:01	31:01	68:01	68:01	31:01	31:01	68:01
	B	13:01	51:01	13:01	51:01	13:01	51:01	13:01	51:01
	C	04:03	14:02	04:03	14:02	04:03	14:02	04:03	14:02
	DRB1	13:01	15:01	13:01	15:01	13:01	15:01	13:01	15:01
Пациент 6	A	24:02	68:68	24:02	68:68	24:02	68:68	24:02	68:68
	B	07:02	40:02	07:02	40:02	07:02	40:0	07:02	40:02

Образец	Гены	Молекулярное типирование		центры секвенирования					
				Macrogen		Peterhof		Illumina	
	C	07:01	15:02	07:01	15:02	07:01	15:02	07:01	15:02
	DRB1	04:05	15:03	04:05	15:03	04:05	15:03	04:05	15:03
Пациент 7	A	02:01	26:01	02:01	26:01				
	B	08:01	39:01	08:01	39:01				
	C	07:01	12:03	07:01	12:03				
	DRB1	03:01	13:01	03:01	13:01				

Таблица 2. Результаты генотипирования по аллелям генов ГКГС, полученные при помощи программ HLA-LA + bowtie2. Для каждого образца даны генотипы аллелей генов ГКГС, полученные в каждой повторности (с помощью молекулярного типирования или из полногеномного секвенирования в трех массивах данных. Красным выделены аллели, отличающиеся от молекулярного типирования.